**Survival Prediction on Titanic Dataset**

**Aim:**

Understand and clean the data. Select relevant features from the training dataset to train a machine learning model to predict survivors (0 or 1) on the test test.

**Dataset:**

The train dataset consists of 891 rows of passenger data.

418 passenger data in test set.

Columns: Passenger ID, Name, Sex, Age, Pclass (3 classes), SibSp, Parch, Ticket, Fare, Cabin, Embarked (3 areas), Survived (0 or 1).

**Initial Analysis/ Understanding the train dataset:**

Women and Child first policy is followed during Evacuation. Hence Sex and Age does play a role in predicting the survived on test data.

1. **Data Type:**

PassengerId int64

Survived int64

Pclass int64

Name object

Sex object

Age float64

SibSp int64

Parch int64

Ticket object

Fare float64

Cabin object

Embarked object

1. **Number of NULL/Missing values:**

Age: 177

Cabin: 687

Embarked: 2

1. **Identifying features:**

Categorical Features: Sex, Pclass, Embarked

Continuous Features: Age

Discrete Features: SibSp, Parch

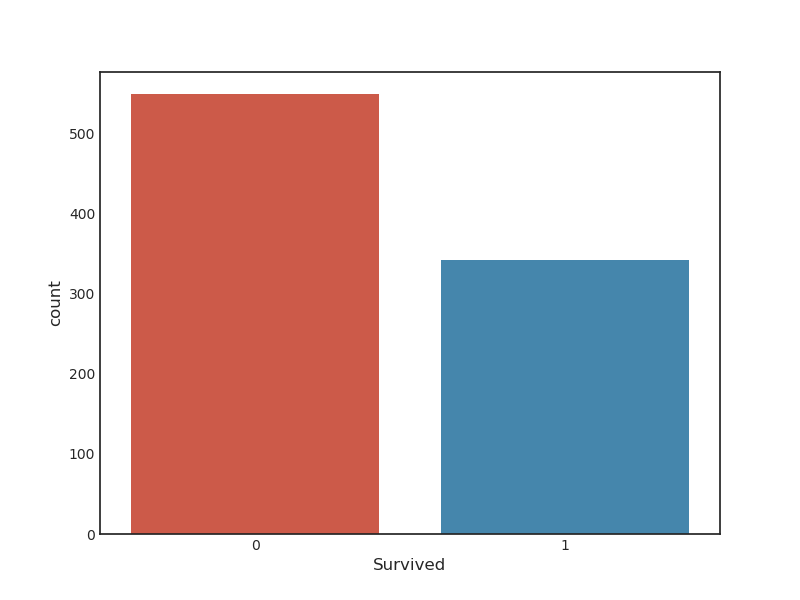
1. **Features not required:**

Passenger ID : not useful

Cabin: cabin numbers are not required as Pclass explains the same in categories.

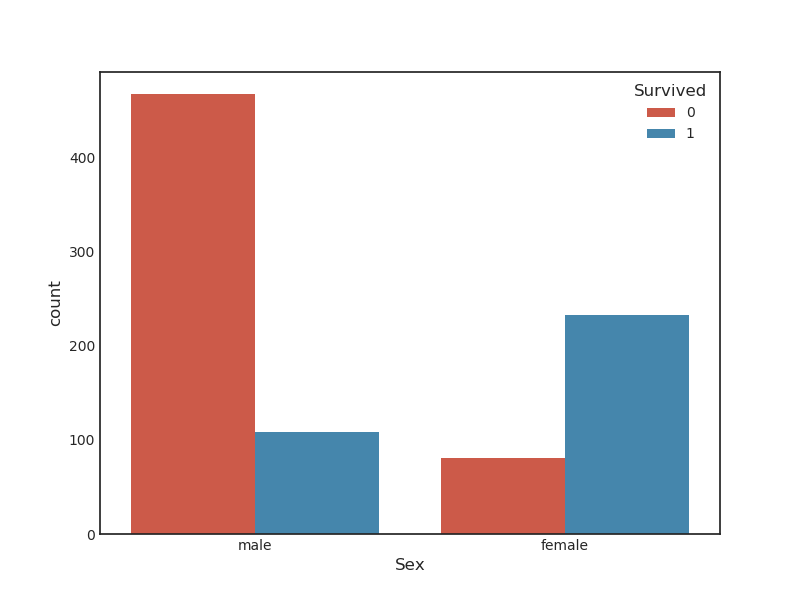
Ticket: Ticket number is not useful

Fare: Fare is also associated with Pclass. Higher the class, higher the fare. Hence not required.

**Exploratory Data Analysis (EDA):**

1. **Survived:**

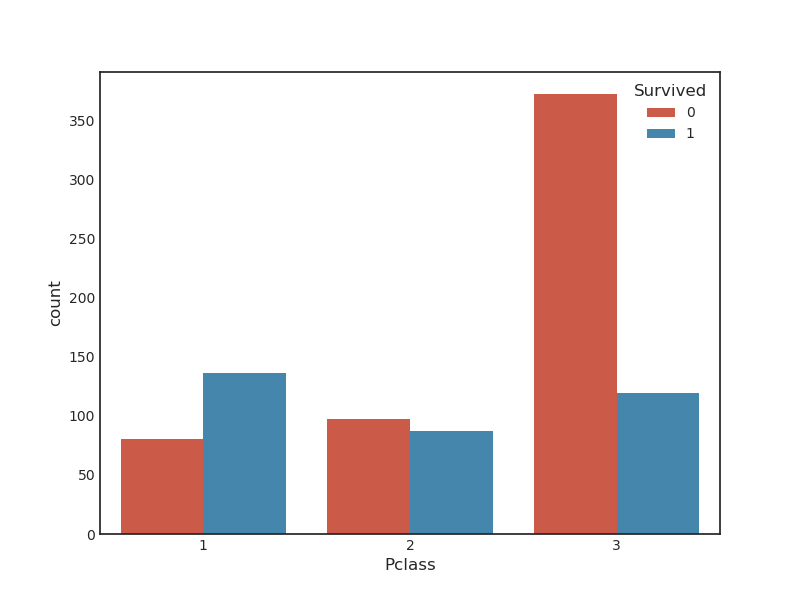
Only 342 people out of 891 survived.

****

**2. Sex Vs Survived:**

Females seem to suvive more than the men, even though men are more than women in number.

This strengthens the “Women and Children First” policy.

**3. Pclass Vs Survived:**

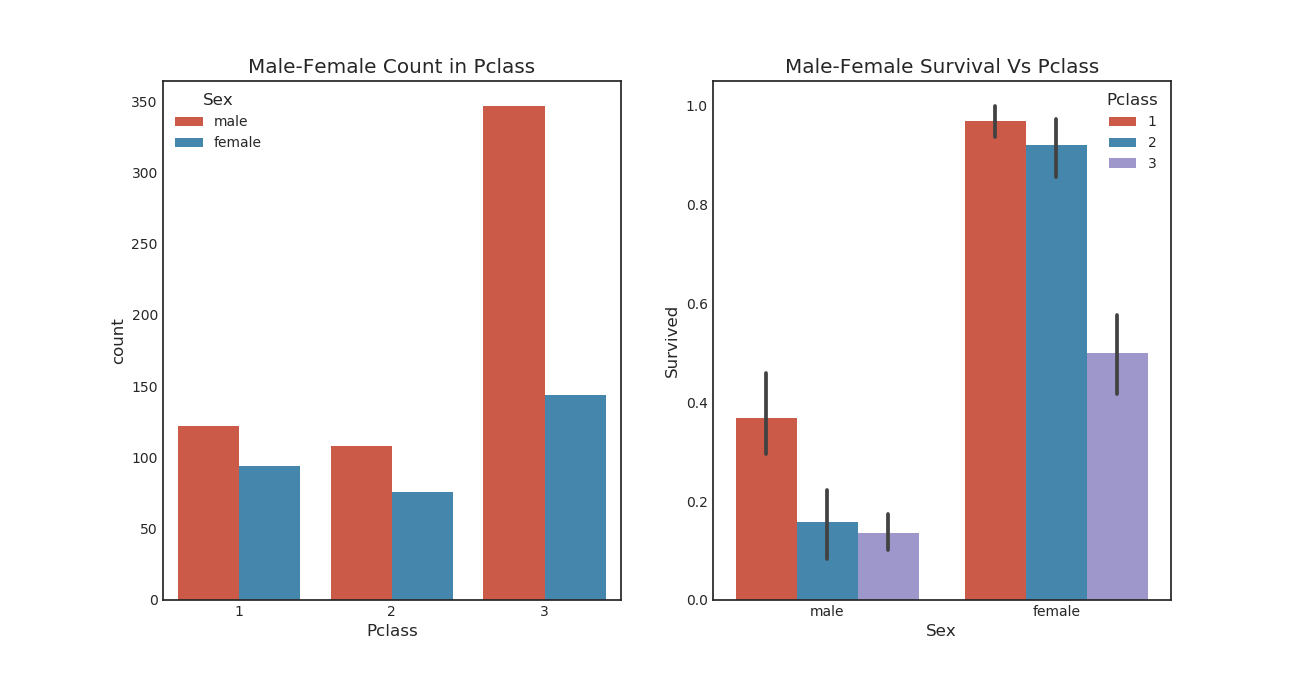
We can see that people belong to 1st class were given high priority. Even though there are many people in 3rd class, only few survived.

Hence, Pclass pose as an important feature.

**4. Pclass Vs Sex – Survived:**

Need to find out if women in class-3 were given equal priority as 1st class following the Policy.

We can see that 1st class is given priority in both sexes, but overall the women were given priority over men regardless of their class.

**5. Age :**

We have **missing values** in this column.

Maximum age = 80

Minimum age = 0.42

Mean of age = 29.69

**Filling missing values:**

There are methods like mean, median and mode to fill the missing values.

But remember “Women and Child first priority” policy. We cannot fill in average age of 29 to a child.

**Method:** Extract titles from the name. Assign means of respective titles to their respective missing values.

Sex / female male

Title

Capt 0 1

Col 0 2

Countess 1 0

Don 0 1

Dr 1 6

Jonkheer 0 1

Lady 1 0

Major 0 2

Master 0 40

Miss 182 0

Mlle 2 0

Mme 1 0

Mr 0 517

Mrs 125 0

Ms 1 0

Rev 0 6

Sir 0 1

**Mlle and Mme are changed to Miss. Jonkheer, Rev, Capt are categorized as 'Others' (since they might be old). Dr is categorized as Mr since there is only one female and 6 male.**

Sex/ female male

Title

Master 0 40

Miss 186 0

Mr 1 529

Mrs 127 0

Other 0 8

**Get Average Age of each of these Titles**

Title

Master 4.574167

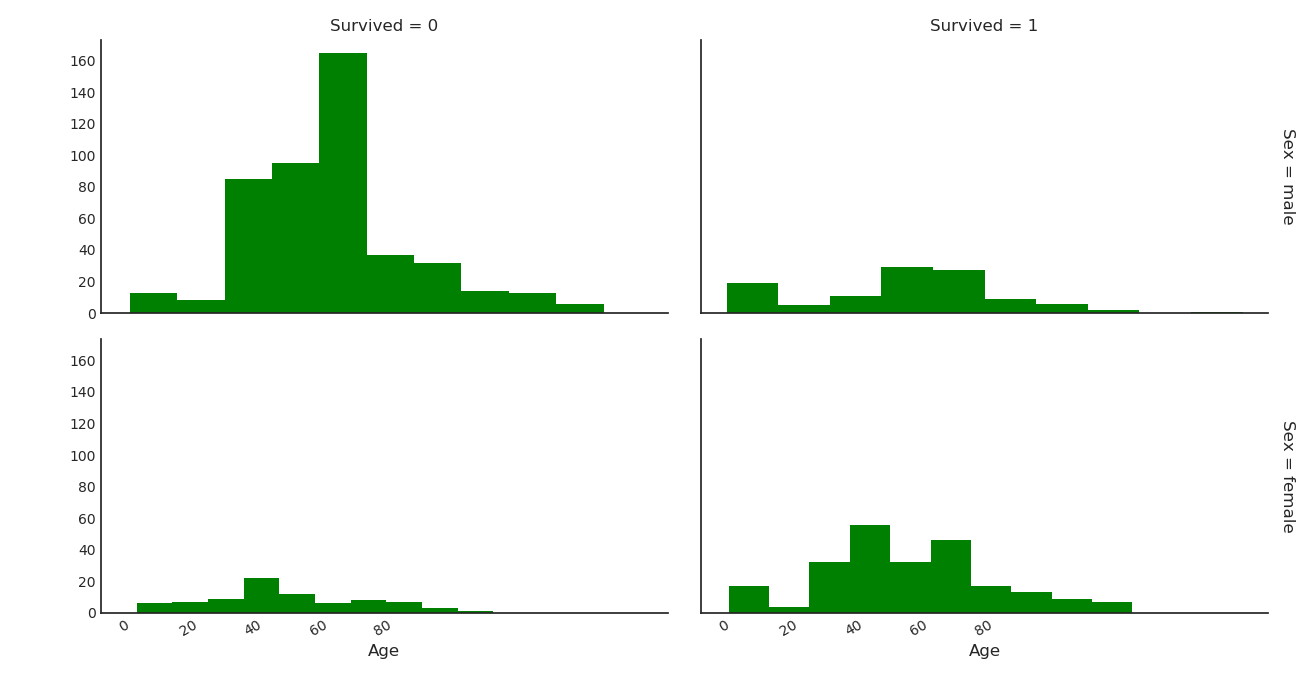
Miss 21.860000

Mr 32.771951

Mrs 35.981818

Other 45.875000

**Fill in the missing values with their respective Title means.**

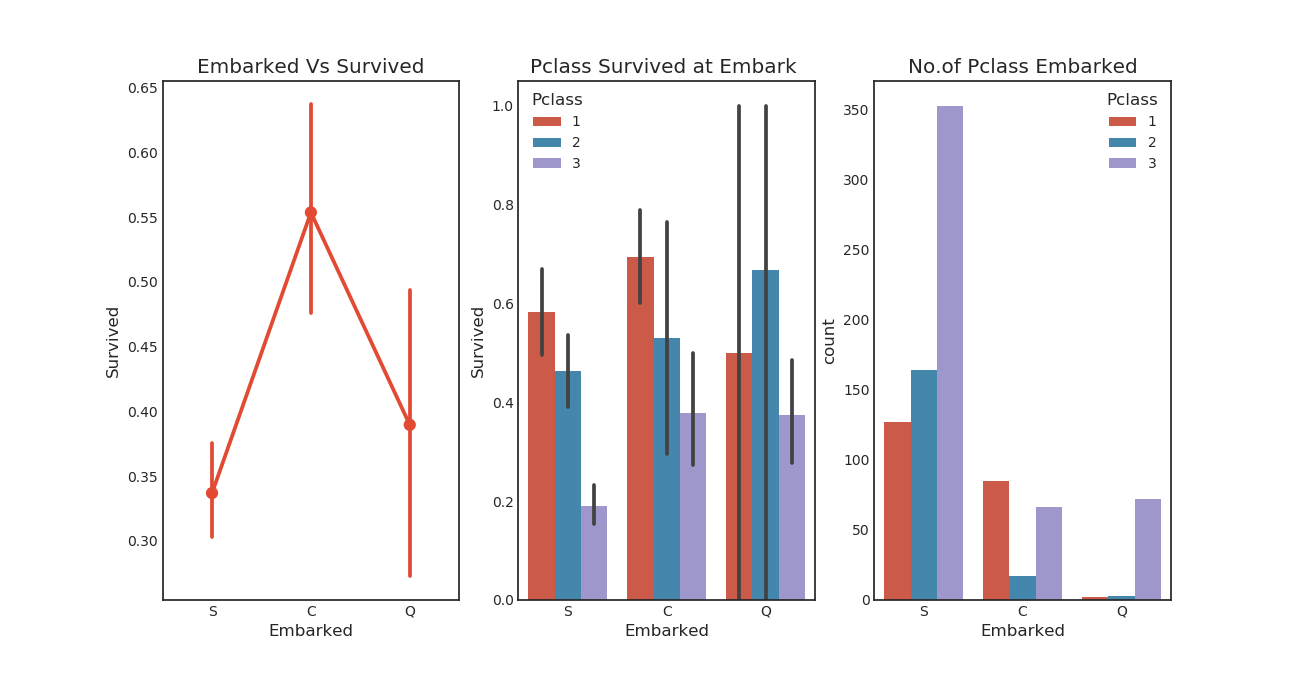
We can see that children and women survived in high number than the male.

So the importance of age and sex holds good.

**6. Embarked:**

It is a categorical value of S, C and Q places. This class has **missing values**.

Lets check the survival of passengers who boarded in these ports.



We can see that people Embarked at port C has survived the most. Pclass 3 has most number boarding at S and are the least survived. We can see that 1st class are given priority irrespective of port boarded.

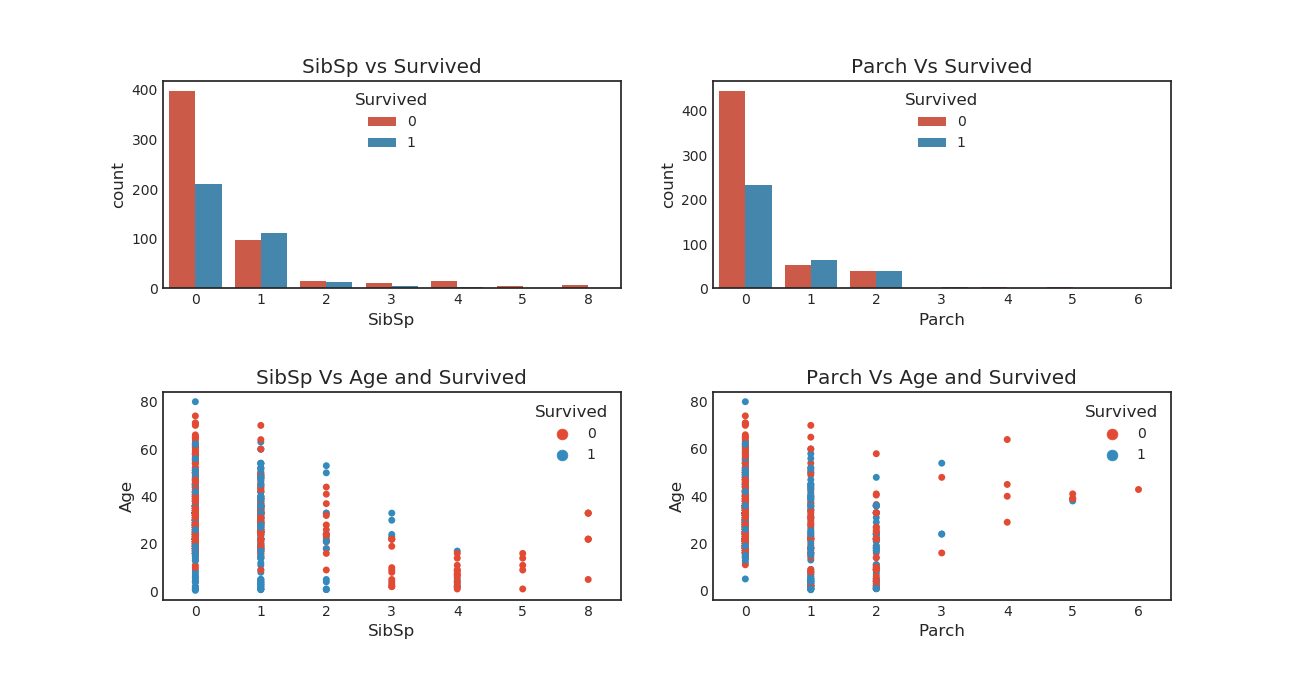
This further strengthens the Pclass feature.

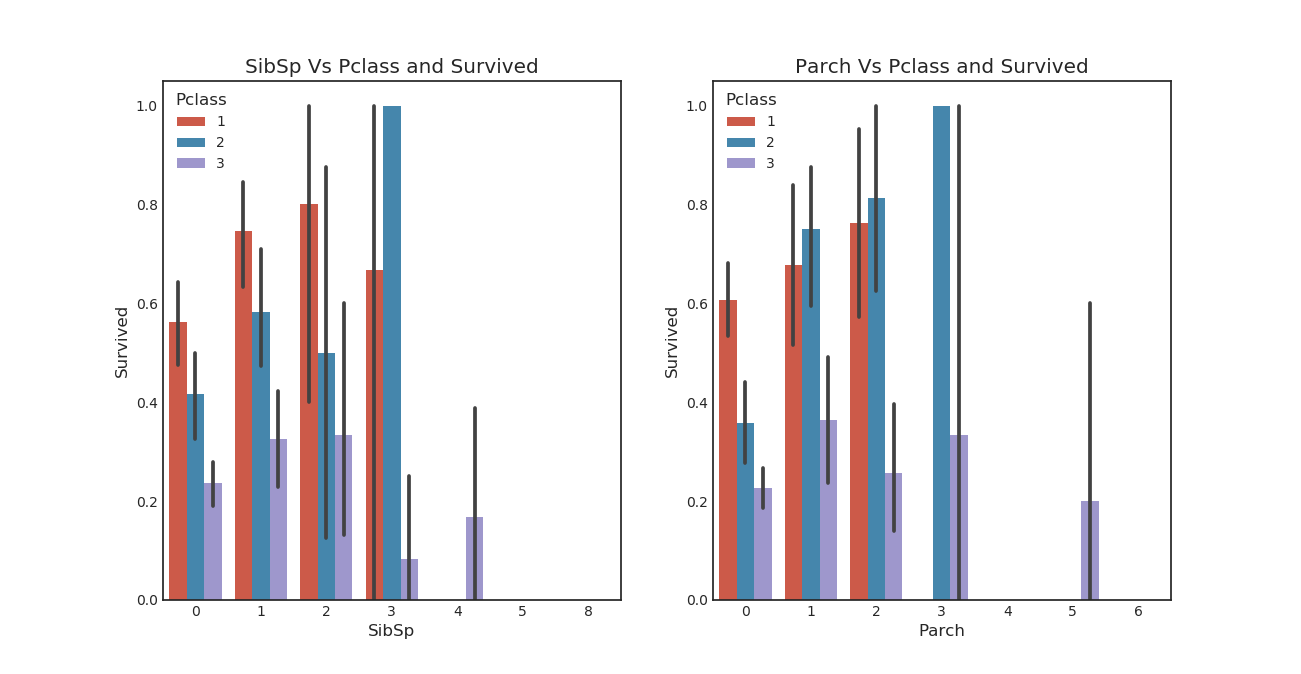
**Filling Null Values:**

**1.** Since maximum of 1st , 2nd and 3rd class passengers boarded at port S, we will fill the NULL values of these classes as S.

2. Check for any NULL values

**7. SibSp /Sibling-Spouse and Parch:**



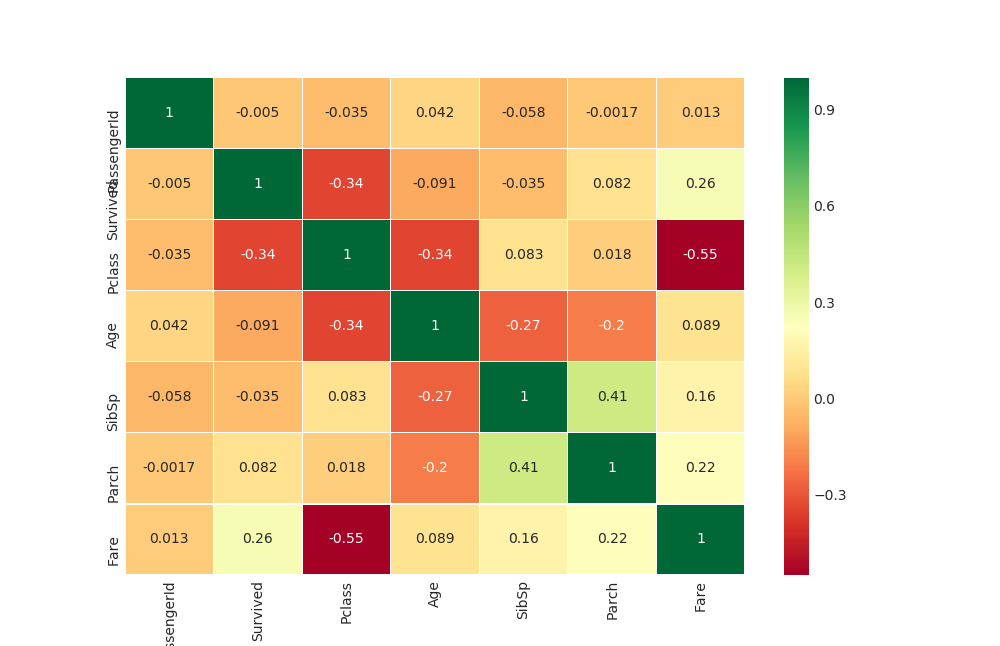


We can see that greater the siblings, lesser the survival. Children were saved regardless of the number of siblings or parents.

We can also see that Pclass also effects the Parch and SibSp.

The chances of survival for passengers with 2-3 siblings/parch seems to be higher.

**Correlation Matrix:**

****

**Final Observations:**

**1.** Passenger\_Id, Fare, Cabin are not useful. Drop these columns

**2.** SibSp and Parch are almost correlated with 0.41. We can drop one of them or combine them to form one feature.

**3.** All other feature variables are not correlated.

**Convert train Data into numerical for Modelling:**

**1. Age:**

Age is continous. Need to change it into categorical.

Min =0, Max = 80. Convert it into bin size of 10. Example Ages 0-10 = 0, Ages 11-20 =1, etc

**2. SibSp and Parch:**

Since both are almost correlated, combine the colums into one – Family.

**3. Sex:**

Convert Sex into 0:Male, 1: Female

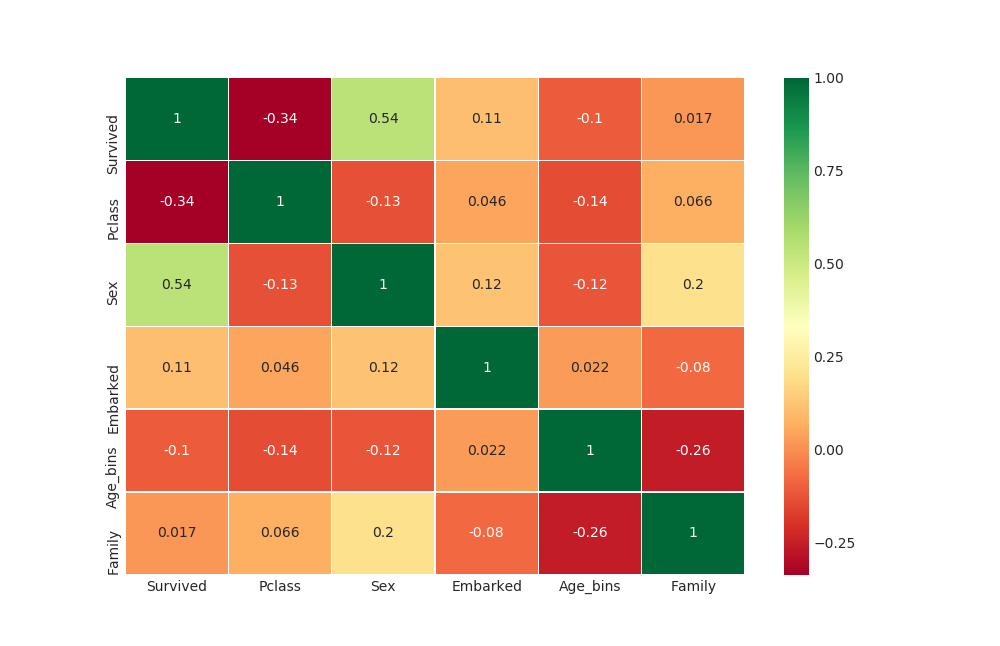
**4. Embark:**

Convert S:0, C:1, Q:2

**Correlation Matrix of the Final Features:**

We can see that there is positive correlation between Sex and Survived.

And all the feature variables are uncorrelated.



**Convert Test Data according to the principals used in train data:**

Using same principles, create Age\_bins, Family features and drop other features.

1. There are 86 NULL Age values and 327 NULL cabin values in test set.

2. Fill Age with logic used in train set.

**Machine learning modelling:**

1. Accuracy for rbf SVM is 0.8246

2.'The accuracy of the KNN is 0.772 for N=5 (default)